ELSEVIER

Analytica Chimica Acta 304 (1995) 257–271

ANALYTICA CHIMICA ACTA

Random error bias in principal component analysis. Part I. derivation of theoretical predictions

N.M. Faber [a,*,1], M.J. Meinders [b], P. Geladi [a], M. Sjöström [a], L.M.C. Buydens [c], G. Kateman [c]

[a] *Research Group for Chemometrics, Umeå University, S-901 87 Umeå, Sweden*
[b] *TNO Centre for Leather and Shoe Research, P.O. Box 135, 5140 AC Waalwijk, Netherlands*
[c] *Department of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands*

Received 1 June 1994; accepted 10 November 1994

Abstract

Principal component analysis (PCA) or singular value decomposition (SVD) are multivariate techniques that are often used to compress large data matrices to a relevant size. Subsequent data analysis then proceeds with the model representation of the data. In this first paper expressions are derived for the prediction of the bias in the eigenvalues of PCA and singular values of SVD that results from random measurement errors in the data. Theoretical expressions for the prediction of this "random error bias" have been given in the statistics literature. These results are, however, restricted to the case that only one principal component (PC) is significant. The first objective of this paper is to extend these results to an arbitrary number of significant PCs. For the generalization Malinowski's error functions are used. A signal-to-noise ratio is defined that describes the error situation for each individual PC. This definition enhances the interpretability of the derived expressions. The adequacy of the derived expressions is tested by a limited Monte Carlo study. This finally leads to the second objective of this paper. Simulation results are always restricted to the class of data that is well represented in the study. Thus rather than giving extensive simulation results it is outlined how the validation and evaluation of theoretical predictions can proceed for a specific application in practice.

Keywords: Chemometrics; Principal component analysis; Singular value decomposition

1. Introduction

Principal component analysis (PCA) or singular value decomposition (SVD) are multivariate techniques that are often used to compress large data matrices in such a way that the relevant information is preserved in the model while most of the noise is discarded [1]. Subsequent data analysis then proceeds with the principal component (PC) model of the data. If quantitative information is to be extracted from the PC model it is essential to analyze the influence of measurement noise. In a previous paper the influence of random measurement errors on the standard errors in PCA was discussed [2]. In this paper the bias (or

* Corresponding author.
[1] Present address: Center for Process Analytical Chemistry, Department of Chemistry, BG-10, University of Washington, Seattle, WA 98195, USA.

SSDI 0003-2670(94)00585-0

systematic error) in PCA that results from random errors in the data is treated. This so-called *random error bias* is essentially different from the bias that results from deleting PCs as in principal component regression (PCR). The idea behind this "biased" regression technique is that deleting the appropriate PCs may yield a better predictive model at the expense of introducing a so-called *underfactoring bias* [3]. In this paper the amount of bias that is already present in the PCs themselves is emphasized.

It is important to note that bias and standard error are very different in nature. The bias is a systematic error, i.e. it leads to a constant offset in the result whereas the standard error is random. This is an unfortunate situation, since, due to its systematic nature, bias can not be quantified by, for example, replicate measurements. Clearly this is not the consequence of this bias being negligible. However, only if one repeatedly finds an unphysical value (e.g. negative concentrations), then one might become alert for possible bias. Thus it is desirable to have an expression by which the size of the offset may be predicted in the same way as the estimate of the true value itself, i.e. directly from the measured data.

The fact that random error bias is so difficult to detect may be the reason why it has seldom been addressed in the (analytical) chemistry literature. Two notable exceptions were found. First there is the early contribution of Moran and Kowalski [4] who work out a detailed theory for predicting the bias for the generalized standard addition method (GSAM) and second there is the very recent paper of Booksh and Kowalski [5] who present simulation results that clearly show the importance of bias for the generalized rank annihilation method (GRAM). It should be noted that bias has been discussed in the statistics literature with respect to multivariate techniques that are frequently used in analytical chemistry. For example, Hodges and Moore [6] give expressions for the bias in ordinary least squares (OLS) resulting from random errors in the independent variables. Furthermore, in a little known paper Goodman and Haberman [7] give theoretical expressions for PCA. Their results are, however, restricted to the case where the data are adequately modelled with only one PC. Since PCA is a cornerstone method for the analysis of multivariate data in any science, it is important to generalize this result.

The first objective of this paper is to show that it is straightforward to extend the statistical result of Goodman and Haberman to an arbitrary number of PCs by borrowing concepts from analytical chemistry. These concepts are the familiar error functions of Malinowski [1]. Although Malinowski's error functions assume uncorrelated and homoscedastic noise, the practical usefulness of the obtained results is not automatically restricted to cases for which this idealized noise model applies, since these simplifying assumptions have an advantage that should not be underestimated: they lead to interpretable results with respect to the sources of error. Expressions for predicting the errors for more complicated noise models are usually more involved but contain the same essential elements. Thus the insight obtained from the expression derived for a simple noise model can often be used in order to efficiently improve the data in the case that a more complicated noise model holds (and it will be cumbersome to develop an error theory). The second objective of this paper is to sketch how the validation and evaluation of theoretical predictions can proceed in practice. In this way we hope to show how future researchers can derive a rule that guides the application of these theoretical predictions in their specific situation. The necessary steps are illustrated by the results of a small Monte Carlo (MC) study.

2. Theory

The organization of this section is as follows. After introducing the notation that aims at emphasizing the different kinds of error (random and systematic), a simple example is worked out which shows why one should expect that *in general* a calculated result is biased as a result of random measurement noise. Next the well known equivalence of PCA and singular value decomposition (SVD) is briefly outlined. In the following part it is discussed which forms of data preprocessing are considered in this paper. A geometric interpretation of the singular values is given and some results for the standard errors in the eigenvalues and singular values are summarized. The geometrical interpretation of the singular values is combined with the standard error in order to obtain a definition for the signal-to-noise

ratio for the individual PCs. Next the expressions for the bias in the eigenvalues and singular values are derived. One advantage of the previously defined signal-to-noise ratio becomes apparent here: the derived bias expressions can be cast in a form that greatly facilitates the interpretation. Finally, the experimental validation and the practical evaluation of the bias expressions are treated.

2.1. Notation

The following notation will be adopted throughout this paper. Bold upper-case letters will denote matrices, e.g. **M**. Bold lower-case letters will denote column vectors, e.g. **u**. Matrix and vector transposition are indicated by a superior "T". $\mathbf{M}^+$ stands for the pseudoinverse of **M**. Italic letters (upper-case as well as lower case) will denote scalars, e.g. X. The locations (row and column) of a matrix element will be specified by indices, e.g. M_{ij} is the element in row i and column j of **M**. The diagonal elements of diagonal matrices, e.g. Λ_{aa} and Θ_{aa}, are denoted by lower case letters with one index symbolizing the position on the diagonal, e.g. λ_a and θ_a.

The following notation is used with respect to measured and errorless quantities. The measured quantity is denoted by adding a "tilde" to the unadorned symbol for the errorless quantity (true value), e.g. $\tilde{X}$. The random error in a measured quantity $\tilde{X}$ (defined for a single realization of the data) is denoted by ϵ_X. Thus we may express a *measured* quantity $\tilde{X}$ in its errorless counterpart and the random error as

$$\tilde{X} = X + \epsilon_X$$

The size of the random error in $\tilde{X}$ is denoted by σ_X, and defined as

$$\sigma_X = \left(E\left[\epsilon_X^2\right]\right)^{1/2}$$

where $E[\bullet]$ denotes the expected value. The expected value is by definition the average for an infinite number of realizations of the data. It is a hypothetical quantity which is, for example, useful for the evaluation of possible bias.

Contrary to the measured quantities the estimated quantities may also have a systematic error or bias. Estimators will be indicated by a "hat", e.g. $\hat{Z}$. A

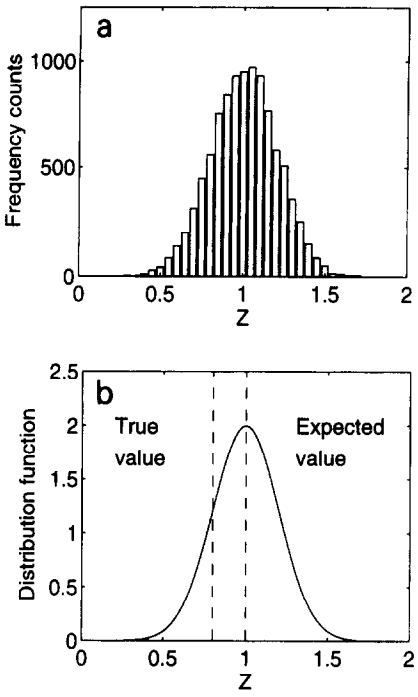

Fig. 1. (a) Histogram and (b) distribution function of the estimated quantity $\hat{Z}$. From the outcome of many realizations of $\hat{Z}$ a histogram can be set up. (The histogram depicted in (a) is constructed from 10000 random numbers drawn from a normal distribution with mean 1 and standard deviation 0.2.) For an infinite number of realizations of $\hat{Z}$ the bins of the histogram can be made arbitrarily small and (after normalization) the distribution function is obtained. (The distribution function in (b) is the normal distribution that has led to the histogram in (a).) The distribution function gives the probability of finding a specific value of $\hat{Z}$. From the distribution function it is clear how the realizations of $\hat{Z}$ cluster around the expected value. The quantity $\hat{Z}$ is called biased if the difference between the average value and the true value does not decrease when the number of realizations of $\hat{Z}$ increases.

quantity is called biased if in the long run (i.e. many realizations) the average will not tend to converge to the true value. This situation is depicted in Fig. 1. The bias in an estimated quantity $\hat{Z}$ is denoted by b_Z and defined as

$$b_Z = E[\hat{Z}] - Z$$

Thus we may express the *estimated* quantity $\hat{Z}$ in its errorless counterpart and the error (random and systematic) as

$$\hat{Z} = Z + \epsilon_Z + b_Z$$

It is seen that the bias can be interpreted as a constant background. In this first part the focus will

be on the derivation of expressions for the prediction of the size of this bias. In the second part the focus will be on discussing the consequences of the presence of random error bias on a number of multivariate problems.

2.2. Bias resulting from random measurement errors

It is important to note that in this paper the bias in an estimated quantity is discussed that results from random (uncorrelated and homoscedastic) measurement errors, i.e. there is no bias in the data. How this random error bias arises and which assumptions have to be made in order to predict it is best explained using a simple example. Assume that one measures a quantity $\tilde{X}$ and wants to estimate the quantity $Z = X^2$. From the preceding section it is known that $\tilde{X}$ may be expressed as

$$\tilde{X} = X + \epsilon_X$$

Then the question of the (possible) bias in the estimated $\hat{Z}$ reduces to the question whether $\hat{Z}$ can be expressed in a similar way as

$$\hat{Z} = Z + \epsilon_Z$$

In order to find out whether a quantity is biased one must evaluate its expectation. Inserting the relevant expressions from the preceding section gives

$$\begin{aligned} E[\hat{Z}] &= E[\tilde{X}^2] \\ &= E\left[(X + \epsilon_X)^2\right] \\ &= E[X^2] + 2XE[\epsilon_X] + E\left[\epsilon_X^2\right] = X^2 + \sigma_X^2 \\ &= Z + \sigma_X^2 \end{aligned}$$

since $E[\epsilon_X] = 0$ for uncorrelated noise. It follows that $\hat{Z}$ is biased ($E[\hat{Z}] \neq Z$), i.e. $\hat{Z}$ should be expressed as

$$\hat{Z} = Z + \epsilon_Z + b_Z$$

and the bias in $\hat{Z}$ is given as

$$b_Z = E[\hat{Z}] - Z = \sigma_X^2$$

Thus the bias in estimating the square of a quantity (from measuring the quantity itself) is equal to the square of the standard deviation of the measurement error.

This should not come as a surprise, since in fact we have just rewritten the equation from which one usually estimates the standard deviation of (homoscedastic and uncorrelated) measurement noise, i.e.

$$\sigma_X^2 = E[\tilde{X}^2] - \left(E[\tilde{X}]\right)^2$$

However, in this paper the attention is focused on bias and this example is merely worked out to show that random errors do not automatically cancel out in an estimated quantity. Random errors will only cancel out if the estimate results from a *linear transformation* of the data.

From the preceding example it should be clear that one does not have to make (strong) assumptions about the distribution of the measurement noise. All that needs to be known with respect to the noise is the size which is quantified by the standard deviation. Finally, it is important to note that in this simple example bias is essentially the result of a *skewed distribution* of the estimated quantity. If, for example, $\tilde{X}$ is normally distributed, then $\hat{Z} = \tilde{X}^2$ is χ^2 distributed (with one degree of freedom). Bias resulting from a skewed distribution is discussed in detail by Moran and Kowalski [4] (see also Fig. 2 in [2]). We will return to this point in Part II when the construction of confidence intervals is discussed.

2.3. Principal component analysis and singular value decomposition

Principal component analysis (PCA) is a multivariate technique that is often used to compress large data matrices in such a way that the relevant information is preserved in the model while most of the noise is discarded [1]. Subsequent data analysis then proceeds with the compressed data.

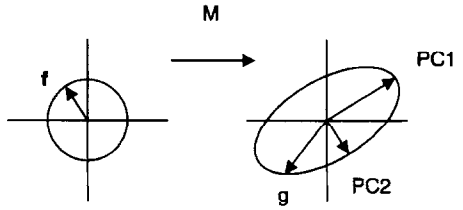

Fig. 2. An $I \times J$ matrix $\mathbf{M}$ can be seen as a linear mapping. A $J \times 1$ vector $\mathbf{f}$ that lies on the unit sphere in $\mathbb{R}^J$ is mapped onto an $I \times 1$ vector $\mathbf{g}$ that lies on an ellipsoid in $\mathbb{R}^I$. The length of the ath semi-axis of this ellipsoid is given by the ath singular value of $\mathbf{M}$, i.e. $\|\mathbf{PC}_a\| = \theta_a$.

PCA consists of two steps. First, one of the cross-product matrices of the $I \times J$ data matrix $\tilde{\mathbf{M}}$, i.e. $\tilde{\mathbf{M}}\tilde{\mathbf{M}}^T$ $(I \times I)$ or $\tilde{\mathbf{M}}^T\tilde{\mathbf{M}}$ $(J \times J)$, is formed. Next, the selected cross-product matrix is eigenanalyzed. Eigenanalysis of $\tilde{\mathbf{M}}\tilde{\mathbf{M}}^T$ and $\tilde{\mathbf{M}}^T\tilde{\mathbf{M}}$ leads to a new vector base of the column and row space of $\tilde{\mathbf{M}}$ respectively. Analysis of the column space is often referred to as Q-mode analysis whereas analysis of the row space is called R-mode analysis. Without loss of generality it is assumed in the remaining part of this paper that $I \geq J$. It is well known that the J non-zero eigenvalues are the same for both modes of analysis:

$$\hat{\lambda}_a = \hat{\mathbf{u}}_a^T(\tilde{\mathbf{M}}\tilde{\mathbf{M}}^T)\hat{\mathbf{u}}_a = \hat{\mathbf{v}}_a^T(\tilde{\mathbf{M}}^T\tilde{\mathbf{M}})\hat{\mathbf{v}}_a \quad \text{for } a = 1,...,J \tag{1}$$

where $\hat{\lambda}_a$ is the ath eigenvalue and $\hat{\mathbf{u}}_a$ and $\hat{\mathbf{v}}_a$ are the associated Q-mode and R-mode eigenvector, respectively (all calculated quantities are estimated). PCA is intimately related to the singular value decomposition (SVD) of $\tilde{\mathbf{M}}$:

$$\hat{\theta}_a = \hat{\mathbf{u}}_a^T\tilde{\mathbf{M}}\hat{\mathbf{v}}_a = \hat{\lambda}_a^{1/2} \quad \text{for } a = 1,...,J \tag{2}$$

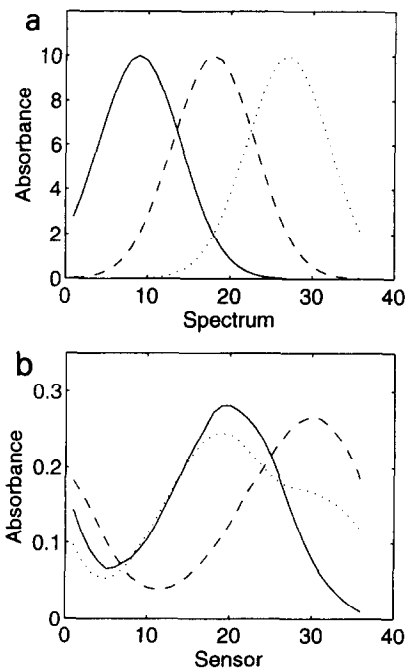

Fig. 3. (a) Denormalized (simulated) HPLC elution profiles and (b) normalized (experimental) UV spectra of adenine (solid line), cytidine (---) and guanine (· · · ·).

where $\hat{\theta}_a$ is the ath singular value. The singular values $\hat{\theta}_a$ are defined to be the positive square roots of the corresponding eigenvalues $\hat{\lambda}_a$. The singular vectors $\hat{\mathbf{u}}_a$ and $\hat{\mathbf{v}}_a$ are normalized and fixed up to pairwise changes of sign. It is seen that all quantities that can be derived from a Q-mode or R-mode PCA are given by the SVD.

According to Malinowski's theory of errors [1] the PCs $(a = 1,...,J)$ can be divided into *primary* $(a = 1,..., A)$ and *secondary* $(a = A + 1,..., J)$ PCs. The primary PCs describe the signal contaminated with measurement error whereas the secondary PCs describe only noise. The number of primary PCs, i.e. A, is called the pseudorank of $\tilde{\mathbf{M}}$. Determining the pseudorank of $\tilde{\mathbf{M}}$ is one of the most challenging problems in multivariate data analysis. The problem of pseudorank estimation is further discussed in Part II with respect to the construction of confidence intervals.

2.4. The influence of data preprocessing

In this paper it is assumed that the data can be adequately described by PCA. However, an adequate model description of the data can only be achieved if the errors are (uncorrelated and) homoscedastic, since PCA is a least-squares fitting technique. If the assumption of homoscedastic errors is violated the data must be scaled (weighted) according to the variance in each data point. Very recently, Paatero and Tapper [8] have discussed a great number of scaling procedures of which balanced scaling gave the best fit of the PCA model to the data. Scaling of the data is of no consequence for the ensuing discussion as long as there exists a "useful" relationship between the factorizations of the raw and the scaled data. See, for example, Eqs. (12)–(19) in [8] that show that the singular values and eigenvalues are not affected in this case. It will therefore be assumed throughout this paper that the data have been scaled appropriately so that the least-squares properties of the PC model hold.

It will be seen that the number of degrees of freedom present in the data plays an important role in the derivation of bias. In this paper we will therefore only take the influence of data preprocessing into account that leads to a loss of degrees of freedom. Degrees of freedom are lost if row, column or grand average are subtracted from the data. The

numbers to be used in these cases are given by Mandel [9].

2.5. Geometrical interpretation of the singular values

The singular values of a (general non-square) matrix have a nice geometrical interpretation. This becomes clear if one interprets the matrix as a linear mapping and applies it on vectors that are chosen in a particular way. Applying an $I \times J$ matrix $\mathbf{M}$ to a $J \times 1$ vector $\mathbf{f}$ that lies on the unit sphere in $\mathbb{R}^J$ gives an $I \times 1$ vector $\mathbf{g}(=\mathbf{Mf})$ that lies on an ellipsoid in $\mathbb{R}^I$ for which the length of the semi-axes is given by the singular values of $\mathbf{M}$. This is illustrated in Fig. 2 for the case where $I = J = 2$. This interpretation will be combined with the standard errors from the next section to define a signal-to-noise ratio for each "principal" direction.

2.6. Standard error in the eigenvalues of PCA and singular values of SVD

Before deriving the expressions for the bias in the eigenvalues $\hat{\lambda}_a$ and singular values $\hat{\theta}_a$ some useful results are summarized for the standard errors in these quantities, i.e. σ_λ and σ_θ. Surely, it is only interesting to look for bias if it makes a significant contribution to the total error. The common measure for the total error in an estimated quantity, say $\hat{Z}$, is the mean square error (*mse*). It is defined as the expected spread around the true value Z:

$$mse_Z = E\left[(\hat{Z} - Z)^2\right]$$

Working out this expression in terms of standard error and bias yields the well known result that

$$mse_Z = \sigma_Z^2 + b_Z^2$$

It is seen that standard error and bias contribute in the same way to the mean square error and a useful measure for the relative importance of the bias should thus be given by the "relative bias", *rb*, defined by

$$rb_Z = b_Z / \sigma_Z$$

This motivates reviewing the expressions for the standard errors.

In another paper [2] it was shown that the standard error in the eigenvalue $\hat{\lambda}_a$ can be predicted as

$$\sigma_{\lambda_a} = 2\lambda_a^{1/2}\sigma_M \quad \text{for } a = 1,...,J \tag{3}$$

where σ_M denotes the true error in the data matrix which has to be estimated. This may be considered to be the main difficulty with such expressions. (The true eigenvalue in Eq. (3) is simply replaced by the estimated value $\hat{\lambda}_a$.) It was shown that the resulting predictions are very accurate if the eigenvalue under consideration is well-separated from the remaining eigenvalues. For the secondary PCs the prediction by Eq. (3) may constitute an overestimate by a factor two [2]. This result is quite satisfactory, since one is only interested in describing the primary PCs. The predicted standard error in the singular value $\hat{\theta}_a$ is given by [2,7]

$$\sigma_{\theta_a} = \sigma_M \quad \text{for } a = 1,...,J \tag{4}$$

It is seen that the predicted standard error in the singular values is constant and equal to the size of the measurement error.

2.7. Signal-to-noise ratio for the individual PCs

In a preceding section it was shown that the singular values have a simple geometrical interpretation: they represent the length of the principal axes. In general, however, the significant singular values may differ orders of magnitude. It is therefore not to be expected that bias will affect all PCs in an identical way. Thus it should be useful to have a measure that describes the error situation for each individual PC.

In analytical chemistry the common measure is the signal-to-noise ratio, ρ. Several definitions of ρ for a data matrix can be found in the literature (see e.g. [10,11]). These definitions usually lead to a ρ for the individual matrix elements (see example below). From the preceding argument it should be clear that an element oriented definition of ρ will not satisfy our demands, since here the focus is on the individual PCs. A logical definition of ρ for the

current problem seems to be the ratio of the singular value and its associated error given by Eq. (4):

$$\rho_a = \theta_a / \sigma_M$$

$$\text{for } a = 1,..., A \qquad (5)$$

Since in the current definition one looks at the signal contribution of a PC to all matrix elements, one expects to find a relatively high value for ρ compared to a value following from an element oriented approach. This may be illustrated by the following example. In another paper [12] a simple 20×10 "experimental" data matrix was simulated by adding normally distributed noise with $\sigma_M = 1$ to constant elements of size M. A straightforward element oriented definition of the signal-to-noise ratio, in the sequel denoted by ρ^*, immediately results in $\rho^* = M$. (Since we have only one primary PC, every reasonable element oriented definition should give the same value.) It was found that for $\rho^* \geq 0.5$ theoretical predictions (bias and standard error) work well for this data set. It is easily verified that this value of ρ^* corresponds to a value for $\rho = 0.5\sqrt{200} \approx 7$. The numerical value for ρ is reasonable, since it indicates that the first PC is above the limit of detection. (This data matrix will be further discussed in Part II with respect to pseudorank estimation.)

In essence, the large difference in size between ρ and ρ^* illustrates the powerful noise averaging capability of PCA. The current definition is obviously tailored with respect to the specific application. However, the underlying principle can be generalized to methods that model the data matrix in a different way, e.g. PLS. This would facilitate a comparison between related methods with respect to their sensitivity to random noise in terms of a typically analytical chemical concept. The current definition of the signal-to-noise ratio seems appropriate, since it is not only problem oriented but also flexible enough to enable such a comparison. A geometrical interpretation of PLS as well as standard errors have been given by Phatak et al. [13,14].

2.8. Bias in the eigenvalues of PCA and singular values of SVD

Goodman and Haberman [7] have given expressions for the case that only one PC is significant. (Furthermore, the row, column and grand average are subtracted from the data.) Using Malinowski's error functions their result is easily extended to an arbitrary number of PCs. The expressions are worked out for the case that no degrees of freedom are lost as a result of data preprocessing and it is shown that it is straightforward to also take this effect into account.

First the real error (RE) function [1] is rewritten using an alternative number of degrees of freedom as

$$(RE)_A^2 = \sum_{a=A+1}^{J} \frac{\hat{\lambda}_a}{(I-A)(J-A)} \qquad (6)$$

The RE function provides an adequate estimate for σ_M if it is evaluated for the correct dimensionality, i.e. the pseudorank A. (It is noted that A is a constant in this paper.) The number of degrees of freedom in the denominator is different from the number of degrees of freedom given by Malinowski. However, several authors have found that it gives a better estimate [7–9,15]. There is no conflict, since the number of degrees of freedom is not derived by Malinowski: the essence of Malinowski's error theory is the introduction of two other functions that show how the measurement error is partitioned between the primary and the secondary PCs. Since the measurement error contributes to the *sum of squares* explained by the primary PCs, one expects from the simple example worked out in a previous section that the calculated eigenvalues and singular values are biased. Furthermore, it is expected that Malinowski's error functions should be particularly useful for deriving this bias.

The first function is the extracted error (XE) function. It gives the part of the error that is not built in a model reconstructed data point and is defined by

$$(XE)_A^2 = \sum_{a=A+1}^{J} \frac{\hat{\lambda}_a}{IJ} \qquad (7)$$

The second function is the imbedded error (IE) function. It gives the part of the noise that contributes to the variation in a model reconstructed data point. Inserting $(RE)_A = \hat{\sigma}_M$ and combining Eqs. (6) and (7) yields

$$(IE)_A^2 = (RE)_A^2 - (XE)_A^2 = \frac{A(I+J-A)\hat{\sigma}_M^2}{IJ} \qquad (8)$$

The total variation of the model is given by the sum of the primary eigenvalues. This sum can be written as the sum of the true eigenvalues and the contribution of the imbedded error:

$$\sum_{a=1}^{A} \hat{\lambda}_a = \sum_{a=1}^{A} \lambda_a + IJ(IE)_A^2$$

$$= \sum_{a=1}^{A} \lambda_a + A(I+J-A)\hat{\sigma}_M^2 \qquad (9)$$

The second term in this sum constitutes the total bias in the A significant PCs. If the measurement error is purely random this total bias should be equally dispersed over the individual eigenvalues. Thus the bias in eigenvalue $\hat{\lambda}_a$ is found as

$$b_{\lambda_a} = E[\hat{\lambda}_a] - \lambda_a = (I+J-A)\sigma_M^2$$

$$\text{for } a = 1,...,A \qquad (10)$$

Here, the substitution $E[\hat{\sigma}_M^2] = \sigma_M^2$ has been made. This simplification will be discussed at length with respect to the experimental validation of Eq. (10).

It is seen that the bias is a product of two factors. The first factor is primarily determined by the size of the matrix (usually $I+J \gg A$). Thus, contrary to intuition, collecting more data will not decrease but actually increase the bias. The second factor gives the dependence on the size of the noise. (It should be noted that it is the same dependence as encountered in the simple example worked out earlier.) From Eq. (10) it is immediately clear that reducing the size of the noise is the best way to avoid significant bias.

As stated earlier, a better understanding of the importance of the bias may be provided by the relative bias,

$$rb_{\lambda_a} = \frac{I+J-A}{2}\frac{\sigma_M}{\theta_a} \cong \frac{I+J}{2}\rho_a^{-1}$$

$$\text{for } a = 1,...,A \qquad (11)$$

where the previously defined signal-to-noise ratio ρ has been introduced. Furthermore, it has been assumed for convenience that $I+J \gg A$. It is immediately clear from the far hand side of Eq. (11) that the bias in a particular eigenvalue starts to dominate the standard error if the associated signal-to-noise ratio is smaller than the "average size" of the matrix, $(I+J)/2$. Since many modern instruments yield huge data matrices, this should be an important result. The fact that the expression for the relative bias can be simplified by inserting ρ can be seen as an additional advantage of the current definition for ρ. It can, however, not be assumed that for more complicated latent variable methods like PLS this additional advantage is kept. In that case the analogy of ρ should only be expected to have a purely descriptive value.

The derivation of bias proceeds along the same line if data preprocessing is applied. Table 1 gives the number of degrees of freedom (NDF) to be employed in the derivation and the resulting bias expressions for three cases that are often encountered in practice. It should be mentioned that an additional approximation has to be made in order to obtain straightforward expressions, since the number of degrees of freedom associated to the XE is no longer equal to the number of data points. (Consequently

Table 1
The influence of data preprocessing on the bias in the eigenvalues of PCA

Data preprocessing	NDF for RE	NDF for XE	b_λ
None	$(I-A)(J-A)$	$I \times J$	$(I+J-A)\sigma_M^2$
Column average subtracted [a]	$(I-A-1)(J-A)$	$(I-1)J$	$(I+J-A-1)\sigma_M^2$
Row, column and grand average subtracted [b]	$(I-A-1)(J-A-1)$	$(I-1)(J-1)$	$(I+J-A-2)\sigma_M^2$

[a] Additional approximation $I \gg 1$.
[b] Additional approximation $I \gg 1$ and $J \gg 1$. The resulting bias expression generalizes the result given by Goodman and Haberman for $A = 1$.

there will not be a factor $I \times J$ in Eq. (8) that cancels out exactly after inserting Eq. (8) into Eq. (9).) For large data matrices this approximation (leading to a tendency to overestimate) is completely justified. For example, if only the column average is subtracted the additional approximation error is of order $1/I$. It is important to note that the result given for the case that row, column and grand average are subtracted generalizes the expression of Goodman and Haberman for $A = 1$. Finally, the bias for other combinations of correcting for row, column and grand average should be derived without difficulty.

The bias in the singular values is easily obtained by working out

$$E\left[\left(\theta_a + \epsilon_{\theta_a} + b_{\theta_a}\right)^2\right] = E\left[\lambda_a + \epsilon_a + b_{\lambda_a}\right]$$

$$\text{for } a = 1,\ldots,A \qquad (12)$$

and introducing the derived bias in the eigenvalues:

$$b_{\theta_a} = E\left[\hat{\theta}_a\right] - \theta_a$$

$$= \left[\theta_a^2 + (I + J - A - 1)\sigma_M^2\right]^{1/2} - \theta_a$$

$$\text{for } a = 1,\ldots,A \qquad (13)$$

Finally, assuming that the second term in square brackets is small results in

$$b_{\theta_a} = 1/2(I + J - A - 1)\sigma_M^2/\theta_a$$

$$\text{for } a = 1,\ldots,A \qquad (14)$$

As a result of this approximation the predicted bias in the singular values tends to be relatively high compared to the predicted bias in the eigenvalues. Since the approximation becomes important in the most interesting case, i.e. the situation where bias is *not negligible* compared to the singular value, a thorough evaluation of its consequences is necessary. We will return to this question in the Results and discussion section. In the ensuing discussions we will only refer to Eq. (14) in order to parallel the treatment of Goodman and Haberman. (It is easily verified that the resulting expression for the relative bias in the singular values is almost identical to Eq. (11) and it will therefore not be repeated here.)

There is, however, an essential difference between Eqs. (14) and (10). This difference lies in the presence of the true singular value θ_a in Eq. (14). Consequently, the reduction of bias in the singular values by reducing the size of the matrix is only possible if less influential rows or columns are discarded. This is a sensible result and it should be contrasted to the corresponding result for the eigenvalues where all rows and columns are equally "good" or "bad" with respect to their contribution to bias. It follows that Eq. (10) is not so informative as Eq. (11). Finally, by comparing Eqs. (10) and (14) it should be straightforward to deduce the bias in the singular values from the corresponding bias in the eigenvalues, given in Table 1.

2.9. Experimental validation of bias expressions

The derived bias expressions (10) and (14) are not exact. (This is a distinct difference with the simple illustrative example.) Several approximations have been made in order to obtain these results. Most of these approximations could easily have been avoided or quantified. However, even without these approximations the resulting expressions would not be exact, since an implicit approximation has been made that cannot be avoided or quantified. This approximation concerns Eq. (6) that *implies* that summing the secondary (noise) eigenvalues leads to an *unbiased* estimate of the measurement error σ_M.

It is, however, well known in the statistics literature [7,9] that the size of the secondary eigenvalues is influenced by the size of the primary eigenvalues and so is the estimate of σ_M. This can be illustrated as follows. Assume, for example, that one has a two-component mixture for which one of the components gives a signal contribution to the data that is "buried" in the noise. Then, obviously, evaluating the *RE* function for the "correct" number of PCs, i.e. two, will give an estimate for σ_M that is more "off-target" than if one evaluates the *RE* function with only one PC. However, if the signal of the weak component is increased eventually a point will be reached where one has to add a PC to the model in order to improve the estimate of σ_M. Goodman and Haberman prove that the estimate provided by Eq. (6) underestimates σ_M. (The size of the approximation error is not known in closed form.) As a result the imbedded error will be larger than assumed in Eqs. (8) and (9) and consequently the *real bias* will be underestimated by the *predicted bias*. This effect will be notable if the signal-to-noise ratio is small,

since then one expects to run into the previously described problem. In summary, for data with small signal-to-noise ratio one expects a large bias (since the imbedded error is large) that will be underestimated to an extend that can not be predicted. Thus one needs to validate the derived expressions for their practical usefulness.

The validation of theoretical predictions is usually performed by so-called Monte Carlo (MC) simulations, which are in fact "computer experiments". In MC simulations the errorless data matrix is disturbed by noise in order to mimic the real life situation. One independent realization of "experimental" data is called a MC trial and a number of MC trials is called a MC sample. In the current investigation the eigenvalues and singular values for a large MC sample are evaluated and averaged. Comparing these averages with the quantities for the errorless data gives an estimate for the real bias. The precision of this estimate is given by the standard error in the average which is calculated from the estimated standard error for one trial divided by the square root of the size of the MC sample. (The standard error for one trial is simply derived from the spread within the MC sample.) Since with MC simulations the errorless data are available, these simulations are extremely suitable for demonstrating the existence of bias. One should, however, be careful that the perturbation of the errorless data proceeds according to a realistic noise model (see Booksh and Kowalski [5] for more details).

With MC simulations it is possible to change the "experimental" conditions in such a way that different *regimes* for the signal-to-noise ratio can be distinguished. The direct relation (11) between the signal-to-noise ratio, ρ, and the relative bias, rb, suggests splitting of the total range of ρ into three regimes, depending on whether $rb \ll 1$, $rb \approx 1$ or $rb \gg 1$. One expects that bias will be negligible for "high" values of ρ, i.e. $rb \ll 1$, since then the eigenvalue and singular value under consideration are large. For "intermediate" values of ρ, i.e. $rb \approx 1$, one expects that bias will become important compared to the standard error and hopes that the prediction will work. For "low" values of ρ, i.e. $rb \gg 1$, one expects that bias will overwhelm the standard error and that the accuracy of the prediction will start to break down. At this point the value of ρ can be used to formulate a practical rule stating that the predicted results can no longer be trusted. The validity of such a rule in another situation should, however, be questioned, since even simulations are limited with respect to the complexity of the data that is used. Thus future researchers should always be cautious to use a rule from the literature and test its validity or set up a rule for the specific application at hand, possibly along the line described in this paper. The inherent limited character of an experimental validation is the reason why in this paper the focus is on methodology and only a small simulation study is performed in order to test the adequacy of the derived expressions.

2.10. Practical evaluation of bias expressions

In the following discussion it is assumed that the signal-to-noise ratio is in a regime for which the bias expressions work. The practical evaluation of the derived expressions presents some obvious difficulties, since these expressions contain unknown parameters that have to be replaced by estimates. First there is the pseudorank of the matrix, A, second there is the standard deviation of the measurement noise, σ_M, and finally there is the true singular value, θ_a, in Eq. (14). From Eq. (6) one knows that estimating A and σ_M are closely related problems. In fact, if an accurate estimate of A is available from, for example, chemical knowledge about the system, then an estimate for σ_M can be obtained by simply evaluating Eq. (6). Furthermore, if the residuals are normally distributed, then the sum of the noise eigenvalues will be (approximately) χ^2 distributed and the usual confidence interval for σ_M can be constructed [16]. (A useful observation is that the residuals are more normally distributed than the measurement noise [12].) Conversely, if an estimate of σ_M is available, then A can be estimated, just by evaluating the *RE* function for all dimensions. However, it is noted that an accurate estimate of A is often not necessary, since in practice usually $I + J \gg A$. Thus it is only important to have a dependable estimate for σ_M. This estimate could, for example, be obtained from replication of the measurement. An alternative is to measure in a region that contains no systematic variation, i.e. the so-called zero-component region [17]. The presence of θ_a in Eq. (14)

leads to a trivial problem. Replacing θ_a by $\hat{\theta}_a$ gives a bias estimate that is obviously wrong, since $\hat{\theta}_a$ is biased. However, this bias estimate can be used to correct $\hat{\theta}_a$ after which the corrected value for $\hat{\theta}_a$ is again inserted in Eq. (14) yielding an improved bias estimate, and so on. After n iterations one has

$$b^{(n)} = x/\left(\hat{\theta}_a - b^{(n-1)}\right) \quad (15)$$

where $x = 1/2(I + J\text{-}A\text{-}1)\sigma_M^2$ and $b^{(n)}$ and $b^{(n-1)}$ denote the bias estimate after n and n-1 iterations respectively. (The subscripts for the bias estimates have been dropped for simplifying reasons.) The iterations are started with $n = 1$ and $b^{(0)} = 0$. Since $\hat{\theta}_a$ overestimates the true value, the bias estimate is approached from below. At convergence one has $b^{(n)} = b^{(n-1)} = \hat{b}$ and Eq. (15) reduces to a quadratic equation in $\hat{b}$ with roots

$$\hat{b}_{\pm} = \left(\hat{\theta}_a \pm \sqrt{\hat{\theta}_a^2 - 4x}\right)/2 \quad (16)$$

It follows that the negative root gives the correct bias estimate without iterations. In principle this procedure should lead to a bias estimate that is only slightly "wrong". How much the estimate is wrong depends on the size of the random error in the quantities that are inserted into Eq. (16). For example, if the random error in a singular value is +5%, then the estimate of the bias will be too low by approximately 5%. Obviously one cannot expect to predict the bias without error in a practical situation. There will only be a problem if the error introduced by the approximate theory (see preceding section) is not negligible with respect to the error introduced by the experiment.

3. Experimental

In previous investigations [2,18] HPLC-UV data matrices were constructed in order to test the usefulness of theoretical results like Eqs. (10) and (14). The response of a three-component mixture was simulated by multiplying Gaussian functions and the (normalized) UV-spectra of adenine, cytidine and guanine taken from the work of Zscheile et al. [19]. The HPLC elution profiles and UV spectra are shown in Fig. 3. Artificial normally distributed noise with a standard deviation of 0.5 mAU was added. In this

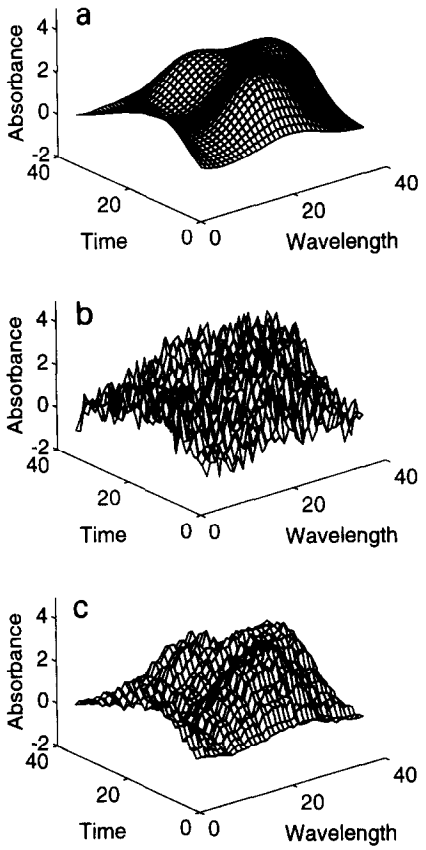

Fig. 4. Simulated HPLC-UV data matrix: (a) before adding noise, (b) after adding noise and (c) after reconstruction by PCA model.

paper we have chosen the heights of the Gaussian functions in such a way that the resulting eigenvalues and singular values range from a value where bias is negligible (and can be predicted very accurately) to a value where the bias overwhelms the standard error (and the prediction is no longer dependable). In this way we hope to demonstrate how other researchers may derive a practical rule that guides the application of these theoretical predictions in their specific situation. For the current simulations the peakheight of adenine, guanine and cytidine is only 10 mAU. Details about the simulated test data are summarized in Table 2.

It should be noted that spectral and chromatographic overlap are severe for this data (for more details, see [2]). Furthermore, the size of the noise encountered for modern UV-diode array detectors (approximately 0.05 mAU) is much smaller than the

Table 2
Characterization of simulated test data [a]

	Adenine	Cytidine	Guanine
Peak positions, μ	9	18	27
Standard deviation peaks, σ	5	5	5
Peakheights, h in mAU	10	10	10
Number of spectra, I		36	
Number of wavelengths, J		36	
σ_M in mAU		0.5	

[a] The elements of the data matrices are generated as $\tilde{M}_{ij} = \Sigma_{k=1}^{K} H_{ik} Y_{jk} + N(0,\sigma_M)$ where K is the number of components (i.e. 3 in our case), H_{ik} is the value of the errorless elution profile of component k at time i, Y_{jk} denotes the errorless absorbance of component k at wavelength j and $N(0,\sigma_M)$ is a normally distributed number with zero mean and standard deviation σ_M. The elements of the elution profiles are calculated as $H_{ik} = h_k \cdot \exp[-1/2(i-\mu_k)^2/\sigma_k^2]$ where the symbols have the meaning as indicated above.

currently simulated value. Thus these data can be claimed to represent a worst case example. A visual impression of the amount of noise present in the data is obtained from Fig. 4 where the simulated data matrix is shown before adding noise, after adding noise and after reconstruction with the three leading PCs. (The noise averaging capability of PCA becomes clear from comparing Figs. 4b and 4c.)

4. Results and discussion

In this section we will confine ourselves to discuss the results obtained for the simulated three-component system without performing data preprocessing, since the effect of data preprocessing is predictable. Furthermore, Eqs. (10) and (14) are evaluated using the errorless quantities. The effect of substituting experimental values for the errorless quantities can be predicted from the size of their standard error as explained in the "practical evaluation" section. In this way it becomes possible to concentrate on the effect of approximation errors that are not so easily quantified.

In Table 3 the results for the primary eigenvalues of PCA are summarized. Since the data matrix is constructed in such a way that there are three detectable components (i.e. $K = 3$) that give a linear and additive contribution to the signal, the pseudo-rank is also three (i.e. $A = 3$). The first column gives the PC under consideration. The second column gives the signal-to-noise ratio, ρ, which is seen to vary over a large range: the ratio of the "average matrix size", i.e. $(I+J)/2 = 36$, and ρ takes the values 0.3, 1.6 and 4.0 respectively. One would therefore expect each PC falling into one of the regimes mentioned in the "experimental validation" section. The third column lists the eigenvalues that are calculated for the errorless data matrix. The fourth column gives the estimated expected value of the eigenvalues in the presence of noise. This estimate is based on a Monte Carlo (MC) sample of 10^6 trials and is seen to be very precise. (The standard errors for one MC trial are estimated to be 62, 12 and 4.5 respectively.) In the next column we give the estimated bias in the eigenvalues, based on the numbers from the two preceding columns. The sixth column lists the predicted values according to Eq. (10). It is seen that the amount of underestimation is *significant* in all cases. The relative error in the predicted bias is given in the last column. The error is approximately the same for the first two PCs and

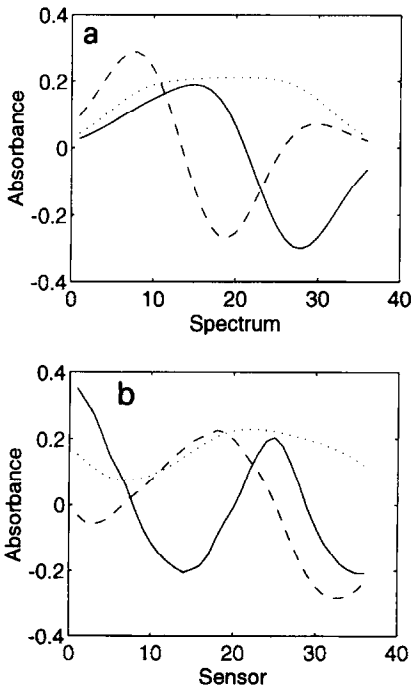

Fig. 5. (a) Abstract elution profiles and (b) abstract spectra for principal component 1 (· · · ·), 2 (- - -) and 3 (solid line) before adding noise.

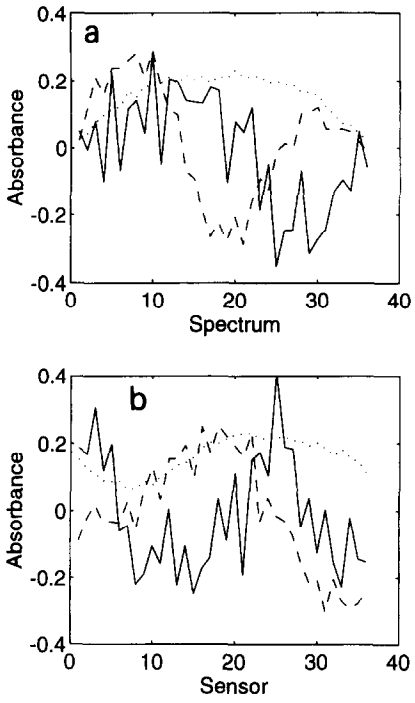

Fig. 6. (a) Abstract elution profiles and (b) abstract spectra for principal component 1 (····), 2 (---) and 3 (solid line) after adding noise.

increases rapidly for the third PC. The error in the theoretical prediction should be compared to the standard error in the eigenvalue for one realization in order to find out whether it is *meaningful*. Then it is immediately clear that the error in the prediction for the first two PCs, i.e. 0.50 and 0.65 respectively, is meaningless (although it is significant), since it is negligible compared to the standard errors, i.e. 62 and 12 respectively. The result for the first PC is obviously best but not interesting, since the bias itself is small compared to the standard error. It is clear that this PC constitutes an example of the first regime for ρ. The result for the second PC, however, is certainly interesting, since here the true bias is larger then the true standard error by a factor 1.5. It is seen that the predicted bias is in excellent agreement with the true value. This PC constitutes an example of desired behaviour in the second regime for ρ. It should be noted that the bias already makes up 14% of the errorless eigenvalue while the bias prediction is still virtually perfect. It is interesting to compare this value with the largest relative bias of 19% reported by Booksh and Kowalski for GRAM [5]. It is reasonable to expect that a theoretical bias prediction will also work well for the amount of bias encountered in that study. As mentioned before, the prediction of the bias in the third eigenvalue is relatively inaccurate. The error in the predicted bias, i.e. $20.55 - 17.25 = 3.3$, is no longer negligible compared to the standard error of 4.5. This PC constitutes an example of breakdown of the theory in the third regime for ρ. However, if one considers that the bias is larger than the errorless eigenvalue itself, then it is justified to qualify this result as promising. One cannot expect accurate quantitative results for a PC that contains such a large amount of imbedded error. This is further illustrated by Figs. 5 and 6 where the abstract elution profiles (the left singular vectors) and abstract spectra (the right singular vectors) are shown before and after adding noise to the data.

From this limited example one might conclude that precise bias estimates should not be expected if ρ is about, say 10. As argued before it will not be possible to derive practical guiding rules that hold in general. Even the attempt is considered to fall beyond the scope of this paper. For example, it is natural to expect that the critical value for ρ depends on the size of the matrix, the number and the relative importance of the significant PCs. This is illustrated

Table 3
Bias in the eigenvalues of PCA

PC	ρ	λ	$E[\hat{\lambda}]$	$E[b_\lambda]$	b_λ	Error in bias
			From simulation		(Eq. 10)	(%)
1	123	3785.06	3802.81(6)	+17.75(6)	+17.25	−2.8
2	22	125.216	143.119(12)	+17.903(12)	+17.25	−3.6
3	9	18.378	38.930(5)	+20.552(5)	+17.25	−16.1

The number in parentheses denotes the standard error in the Monte Carlo sample average (expressed in units of the last reported digit)

Table 4
Bias in the singular values of SVD

PC	ρ	θ	$E[\hat{\theta}]$	$E[b_\theta]$	b_θ	Error in bias
			From simulation		(Eq. 14)	(%)
1	123	61.5228	61.6649(5)	+0.1421(5)	+0.1382	−2.7
2	22	11.1900	11.9535(5)	+0.7635(5)	+0.7596	−0.5
3	9	4.2870	6.2291(4)	+1.9421(4)	+1.9828	+2.1

The number in parentheses denotes the standard error in the Monte Carlo sample average (expressed in units of the last reported digit).

by the results for the 20×10 matrix with constant elements [12]. There it was seen that $\rho \approx 7$ already gives accurate predictions. In order to obtain a reliable critical value for a certain application additional simulations in the expected regimes are necessary taking all the relevant aspects of the data into account.

In Table 4 the results are given for the singular values of SVD. The explanation of the symbols is equivalent to Table 3. It is seen that the relative error in the predicted bias of the first singular value is almost the same as for the corresponding eigenvalue (-2.7% compared to -2.8%). It is, however, considerably smaller for the remaining PCs. This clearly demonstrates the effect of the extra approximation involved in deriving Eq. (14) from Eq. (10). For the last PC the predicted bias even overestimates the real bias. It is, however, not to be expected that this "beneficial" effect will hold in general. The additional approximation leads to a tendency to overestimate but it can not be concluded that it always leads to an effective cancellation of the effect of the implicit approximation with respect to Eq. (6). The data set considered here might just be a favourable case. This further illustrates the fact that simulation results can only be trusted if the data are designed with respect to a specific application.

5. Conclusions

In this paper it is emphasized that bias resulting from random measurement errors should always be expected for an estimated quantity if the calculation involves a non-linear transformation of the measured data. This has been illustrated by the simple example where one measures a quantity in order to estimate its square. There is a strong analogy between this simple illustrative example and PCA where the eigenvalues represent sums of squares. The role of measurement errors in PCA is described by Malinowski's error functions. Using these error functions expressions have been derived that enable the prediction of the bias in the eigenvalues of PCA and singular values of SVD. The derived expressions constitute the generalization of the result given by Goodman and Haberman for the case that only one PC is significant (see Table 1). A definition for the signal-to-noise ratio has been proposed that describes the error situation for each individual PC. An important advantage of the current definition is that it improves the interpretation of the derived bias expressions.

It is shown how the effect of the underlying approximations should be validated and how the expressions should be evaluated in practice. The adequacy of the derived expressions for PCA is validated by a small simulation study. For the data set considered it is shown that depending on the signal-to-noise ratio for a specific PC the derived bias expressions work well. For example, a bias that is 1.5 times the standard error is accurately predicted. It has been argued that simulation results (even for large studies) are always restricted with respect to their usefulness for practical applications. Rather than relying on results presented in the literature future researchers should devise simulations that include data sets that correspond to the complexity of their specific problem. Such a validation could proceed along the line detailed in this study. Finally, it is important to note that the derived bias expressions are based on Malinowski's error functions. Their applicability is therefore – in a strict sense – limited to the case that the measurement noise is uncorrelated and homoscedastic. It is, however, to be expected that these expressions are still useful if the

data matrix is adequately fitted by PCA after the application of balanced scaling [8].

Acknowledgements

NMF acknowledges the financial assistance of a COMETT Ba grant during his stay in Umeå.

References

[1] E.R. Malinowski, Factor Analysis in Chemistry, 2nd edn., Wiley, New York, 1991.
[2] N.M. Faber, L.M.C. Buydens and G. Kateman, J. Chemom., 7 (1993) 495.
[3] J. Mandel, Am. Stat., 36 (1982) 15.
[4] M.G. Moran and B.R. Kowalski, Anal. Chem., 56 (1984) 562.
[5] K. Booksh and B.R. Kowalski, J. Chemom., 8 (1994) 45.
[6] S.D. Hodges and P.G. Moore, Appl. Stat., 21 (1972) 185.
[7] L.A. Goodman and S.J. Haberman, JASA, 85 (1990) 139.
[8] P. Paatero and U. Tapper, Chemom. Intell. Lab. Syst., 18 (1993) 183.
[9] J. Mandel, Technometrics, 13 (1971) 1.
[10] G.W. Stewart, Contemp. Math., 112 (1990) 171.
[11] Y. Wang, O.S. Borgen and B.R. Kowalski, J. Chemom., 7 (1993) 439.
[12] N.M. Faber, L.M.C. Buydens and G. Kateman, Chemom. Intell. Lab. Syst., 25 (1994) 203.
[13] A. Phatak, P.M. Reilly and A. Penlidis, Communn. Stat. Theory Meth., 21 (1992) 1517.
[14] A. Phatak, P.M. Reilly and A. Penlidis, Anal. Chim. Acta, 277 (1993) 495.
[15] E.A. Sylvestre, W.H. Lawton and M.S. Maggio, Technometrics, 16 (1974) 353.
[16] J.R. Green and D. Margerison, Statistical Treatment of Experimental Data, Elsevier, Amsterdam, 1978.
[17] Y.-Z. Liang, O.M. Kvalheim and A. Höskuldsson, J. Chemom., 7 (1993) 277.
[18] N.M. Faber, L.M.C. Buydens and G. Kateman, Anal. Chim. Acta, 296 (1994) 1.
[19] F.P. Zscheile, H.C. Murray, G.A. Baker and R.G. Peddicord, Anal. Chem., 34 (1962) 1776